

# Assessment of automatic ligand building in *ARP/wARP*

Guillaume X. Evrard,<sup>a\*</sup> Gerrit G. Langer,<sup>a</sup> Anastassis Perrakis<sup>b</sup> and Victor S. Lamzin<sup>a</sup>

<sup>a</sup>European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany, and <sup>b</sup>Netherlands Cancer Institute, Department of Molecular Carcinogenesis, Plesmanlaan 21, 1066 CX, Amsterdam, Netherlands

Correspondence e-mail: evrard@embl-hamburg.de

The efficiency of the ligand-building module of *ARP/wARP* version 6.1 has been assessed through extensive tests on a large variety of protein–ligand complexes from the PDB, as available from the Uppsala Electron Density Server. Ligand building in *ARP/wARP* involves two main steps: automatic identification of the location of the ligand and the actual construction of its atomic model. The first step is most successful for large ligands. The second step, ligand construction, is more powerful with X-ray data at high resolution and ligands of small to medium size. Both steps are successful for ligands with low to moderate atomic displacement parameters. The results highlight the strengths and weaknesses of both the method of ligand building and the large-scale validation procedure and help to identify means of further improvement.

Received 2 May 2006  
Accepted 19 June 2006

## 1. Introduction

A key to understanding the function of proteins in their biological context is the modelling and the interpretation of their interactions with small molecules, commonly known as ligands. Several thousand such compounds are known today. Many currently available protein–ligand structures represent the complex of a protein (drug target) with a potential or, less often, an available life-saving drug (Schindler *et al.*, 2000). During lead optimization in the process of structure-assisted drug development, numerous structures need to be determined which essentially represent the same protein complexed with a variety of ligand molecules under different soaking conditions. Although computational modelling and docking studies of large *in silico* libraries play a predominant role in this stage, X-ray crystallography is used as the primary experimental technique to determine protein–ligand complex structures in order to provide detailed structural information.

Development of methods for automatic building of molecular models in X-ray crystallography has so far been focused almost exclusively on the construction of the protein model alone, which was (and often still is) the most labour-intensive and time-consuming step. Less emphasis has been given to the completion of a macromolecular model, which often includes addition and refinement of bound ligand molecules. Particularly in the course of a series of ligand studies, the main goal of automation is to add ligands to an otherwise existing and well refined protein structure. During the last few years, a number of methods have started to emerge that address this topic. There have been advances in bioinformatics for the prediction of a possible ligand and inferring the protein function given a known binding site and its shape (*e.g.* Morris *et al.*, 2005). Crystallographic approaches for building and fitting a ligand of known chemistry to a known or even an unknown location

in the electron density include *BLOB* (Diller *et al.*, 1999), which uses a Monte Carlo technique to position a ligand into an electron-density map in an appropriate conformation, *X-Ligand* (Oldfield, 2001) and *Coot* (Emsley & Cowtan, 2004), which match random conformations of the ligand to the shape of the density cluster through a principal component analysis, a technique utilizing a medial axis transform of an electron-density isosurface (Aishima *et al.*, 2005), *SOLVE/RESOLVE* (Terwilliger *et al.*, 2006), which breaks the ligand into rigid fragments and then links them subject to the density and stereochemical constraints, and *ARP/wARP* (Zwart *et al.*, 2004), which we deal with in this paper in detail.

Like all methods outlined above, the ligand-building module in *ARP/wARP* attempts to fit a complete ligand molecule into electron density. Its main difference from other approaches lies in the construction of a ligand through an iterative atom-by-atom expansion. This may potentially allow the construction of partially occupied ligands and let the researcher know which part of the ligand has tight binding and where to. In order to proceed in this direction, it is essential to extensively analyse the efficiency of the ligand-building module in *ARP/wARP* and to pinpoint its current achievements and limitations. Therefore, as a performance benchmark we attempted the reconstruction of known protein–ligand complexes.

## 2. Methods

The first step of a study that aims at assessing the performance of any automated process is the selection of the test set. To construct the test set here, we chose PDB entries from the Electron Density Server (EDS) in Uppsala (Kleywegt *et al.*, 2004) which contained at least one bound ligand. As implied by the use of the EDS, all these entries have readily available validated associated X-ray diffraction data sets. For convenience in the downstream selection of the test set, these entries were organized in a relational database. The next steps involved automating the execution of the *ARP/wARP* ligand-building procedure over the whole test set and evaluating success or failure on the fly. For each case, this included feeding the program with adequately formatted input (the coordinates of the protein, the diffraction data and a feasible conformer of the relevant ligand with correct stereochemical parameters) and delivering output for post-run analysis. The correctness of the reconstruction was then assessed by comparison of the automatically reconstructed ligand to the original ligand coordinates present in the PDB entry. Finally, an overall analysis was performed to derive indications of software performance and directions for further improvements.

It is obvious that in such an evaluation the final figures depend on each of the preceding steps, *i.e.* on the software itself, on the selected test sample and on the success indicator chosen to judge correct reconstruction. All the numbers obtained in this study are thus relative to the conditions in which the test was conducted. To be of practical interest, the

conditions were set close to those of the ‘standard’ black-box use of *ARP/wARP*.

### 2.1. Algorithm for ligand building

The ligand-building module of *ARP/wARP* version 6.1 (released in July 2004) aims at positioning a bound ligand of known chemistry in a difference electron-density map. The main steps of the algorithm (Zwart *et al.*, 2004) are briefly outlined below.

(i) The protein model without the ligands and solvent is refined with the CCP4 version of *REFMAC5* (Murshudov *et al.*, 1997; Collaborative Computational Project, Number 4, 1994) against the diffraction data. Manual intervention or automatic means of drastically modifying the protein model are not carried out. The resulting structure factors are used to construct a  $\sigma_A$ -weighted difference density map ( $mF_o - DF_c$ ,  $\alpha_c$ ), into which the ligand should be fitted.

(ii) The largest cluster in the difference electron-density map is selected and parameterized using a sparse-grid representation.

(iii) A number of possible ligand models are constructed by iterative assignment of ligand-atom labels to cluster grid nodes.

(iv) All constructed models of the ligand are refined in real space and scores are derived to select the single best model.

A crucial step in this procedure is the volume-based selection of the largest density cluster (step ii). The expected ligand volume is computed from the volume of the ligand multiplied by an excess volume factor whose numerical value can either be the empirical default provided (1.3) or can be adjusted by the user. In the difference density map, a contouring density level  $\rho_{\text{thres}}$  is defined such that the volume of the largest cluster equals the expected ligand volume. Finding a value of  $\rho_{\text{thres}}$  that satisfies this criterion is a simple root-finding problem solved by bracketing and bisection (see, for example, Arfken, 1985). The procedure is thus designed for cases where the ligand to be built is the largest of possibly several ligands bound to the protein. Otherwise, the construction of the ligand may be attempted at a wrong location. Consequently, if there are several ligands present, they can be built one after another, starting with the largest ligand and working through to the smallest. In the current evaluation study, we excluded this consecutive ligand building for the sake of simplicity. This therefore self-defines a strong criterion for discrimination of *a priori* ‘easy cases’, where a large ligand dominates the difference density map, from ‘hard cases’ where the signature of a small ligand may be lost in the presence of larger ligands and even possibly solvent or noise.

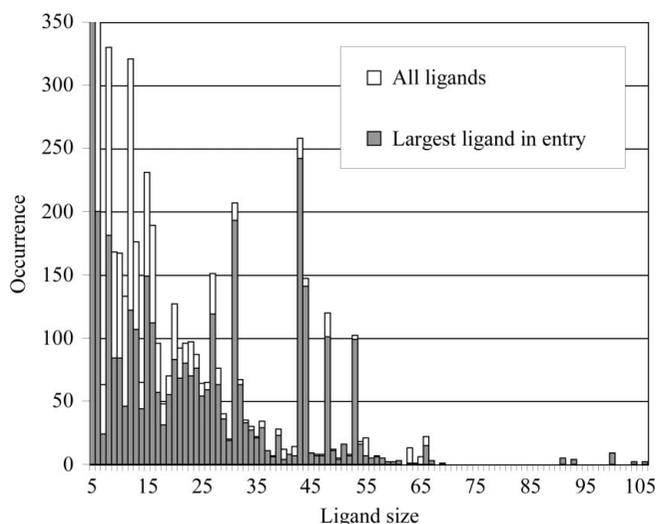
The ligand construction *stricto sensu* (step iii) consists of putting ligand atom labels on the selected density cluster grid nodes (‘label swap’). As described in Zwart *et al.* (2004), on the basis of the stereochemistry of the search model the method defines the order in which atoms are added to a partial ligand model that is iteratively completed. This completion is constrained by the ligand topology and the score for each model takes into account density, bond and angle-bonded

distances, chirality and van der Waals repulsive contributions. The optimization strategy for finding suitable models is a quasi-exhaustive graph search attempting to match the ligand topology to the grid nodes of the density cluster. The fact that the volume of the cluster is set to be somewhat larger than the true volume of the ligand is essential for the construction of many (rather than one) putative models. The amount of search possibilities is defined by a 'branch-and-bound' approach. The intrinsic computational complexity of the ligand construction is mainly dependent on the ligand size and on the resolution of the data. Indeed, the complexity of mapping  $N$  ligand atoms on  $M$  cluster grid nodes increases drastically with an increase in the ligand size. At the same time, low-resolution data and high atomic displacement parameters (ADPs) of the ligand yield a smoothed density cluster where the power of the 'density score' is decreased and the construction of a higher number of putative models becomes necessary.

The final selection of the single 'best' ligand model is carried out at step (iv), when all putative models are subject to stereochemically restrained real-space refinement in the difference electron density. The selected model is that with the lowest goodness-of-fit value.

## 2.2. Defining the test sample

All EDS entries that contain at least one ligand were organized in a simple SQL database with one table for the protein models that included the PDB code, the resolution of the diffraction data, ligand names, occupancy and ADPs of the ligands as well as their real-space  $R$  factors. This table was linked to a second SQL table containing the list of unique ligands with distinct types, their size and topology. Entries suitable for the test sample have been selected by executing database queries with an application of the desired criteria on the fly.



**Figure 1**  
Distribution of ligand sizes (non-H atoms) for the initially selected full sample and the sample of largest ligands. Outstanding peaks reflect the presence of widespread ligands, e.g. HED, ADP, ATP, HEM, NAP, FAD for sizes 8, 27, 31, 43, 48 and 53, respectively.

The large-scale test exemplified the need for thorough pre-processing of the PDB entries. The selected subset of the EDS entries still contained a variety of format errors and inconsistencies. We carried out an additional cleanup and repair of the entries where format-conflicting features could be identified. In addition, the following cases were discarded from the list of building tasks: (i) non-protein cases (DNA or RNA fragments or multi-ligand conglomerates), (ii) cases with DNA or RNA as heterocompounds, (iii) heterocompounds that are part of the protein main chain, (iv) partially built and multimeric ligands and (v) ligands consisting of less than five non-H atoms.

The finally selected test sample consisted of 3884 PDB entries encompassing 1447 unique ligands. From these entries, 6132 ligand-building tasks have been constructed. The number of ligand-building tasks per protein can be larger than one; it is defined by the number of suitable ligand types present, irrespective of the number of their copies.

The *ARP/wARP* version 6.1 ligand-building module (Zwart *et al.*, 2004) has been designed to construct the single largest and fully occupied ligand. Of the selected 6132 tasks, 3369 fulfil the above constraint and hereafter the overall results are presented based on these tasks, with a single exception described in §3.7. We distinguish between 706 small ligands (five or six non-H atoms) and 2663 large ligands.

The test sample is very diverse in terms of the ligand size (Fig. 1), with the largest ligand, adeninylpentylcobalamin in PDB entry 1eex, containing 106 non-H atoms. Ligands with five or six atoms (e.g. sulfate and phosphate ions) largely outnumber other ligands and this was taken into account to avoid bias in the analysis.

## 2.3. Automating the runs

Shell scripts were set up to download the diffraction data (in MTZ format), the PDB files from the EDS and the ligand templates from the Hetero-compound Information Centre in Uppsala (HIC-Up; Kleywegt & Jones, 1998) as well as to automate the preparation and running of ligand-building tasks. We used the 'clean' PDB files for each ligand, which represent ideal coordinates that were prepared using the *CORINA* program (Gasteiger *et al.*, 1990) and are available from the MSD (Golovin *et al.*, 2005). All nonprotein atoms were automatically removed from the PDB files and the *ARP/wARP* ligand-building module was launched. In each building task, the coordinates of the original ligand present in the PDB file (hereafter called the reference ligand) were stored and used for evaluation of the selected ligand site and comparison with the reconstructed ligand. We note that the coordinates of the reference ligands were assumed to be correct.

## 2.4. Assessment of the results

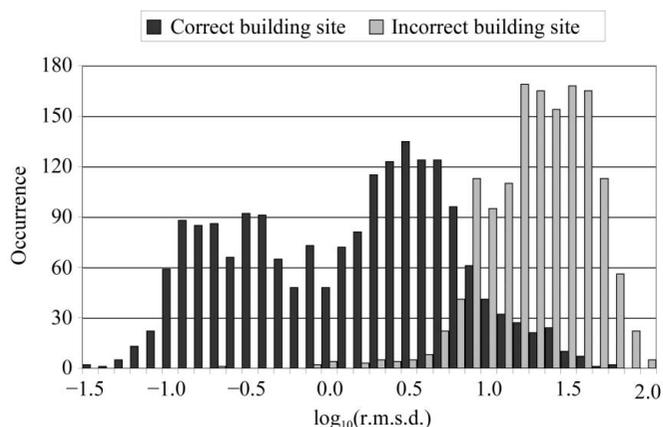
The automatic location of the ligand-binding site was deemed successful if the reference ligand overlapped with the selected density cluster. Overlap was considered to occur if the shortest distance between any atom of the reference ligand and any of the cluster grid points was smaller than 1.0 Å. After

evaluation of the location, assessment of the ligand reconstruction was carried out. This task proved nontrivial since there is no single foolproof measure that indicates success or failure. Ideally, a measure for the correctness of ligand building should take into account both the conformation of the built ligand and the validity of its intrinsic stereochemistry. A simple line-by-line comparison of the PDB files for the reference and built ligand (in other words, computation of the distance for all pairs of atoms with identical atomic labels) is too restrictive and may give false negatives (correct reconstruction reported as being incorrect). For example, possible equivalent conformations of aromatic rings related by a 180° rotation around the C<sup>β</sup>–C<sup>γ</sup> bond in tyrosine or phenylalanine would be considered to be different models. Another measure, the distance between each atom of one ligand and its nearest neighbour in the counterpart, is too loose and often gives a false positive. The metric that performs best in defining a one-to-one mapping is the distance between each atom in the built ligand and its nearest neighbour in the reference ligand, with the constraint that each atom in the latter can be selected only once. The metric is not symmetrical and depends somewhat on the order in which the atoms are taken. However, owing to its inherent nonlinearity it is a robust amplifier of discrepancies between the models and thus acts as a very good quality indicator. The root-mean-square deviation (r.m.s.d.) is subsequently computed as

$$\text{r.m.s.d.} = \left[ \frac{1}{N} \sum_{i=1,N} (\mathbf{x}_i^{\text{ref}} - \mathbf{x}_i^{\text{built}})^2 \right]^{1/2},$$

where  $\mathbf{x}_i^{\text{ref}}$  and  $\mathbf{x}_i^{\text{built}}$  correspond to the coordinate vectors of the  $i$ th atom of the reference and built ligand, respectively;  $N$  is the number of ligand atoms.

The distribution of the r.m.s.d. values obtained from the large-scale test exhibits three main regions (Fig. 2): (i) an interatomic scale (below 1 Å), corresponding to correctly built ligands, (ii) an intramolecular scale (between 1 and 10 Å), where the ligand was built into the correct cluster but in a



**Figure 2** Distribution of r.m.s.d. values between the reconstructed and the reference ligand. The plot highlights the three different regions: 1.0 Å and below, correctly built ligand in correctly identified site; between 1.0 Å and 10 Å, the ligand is at a correct site but not properly built; above 10 Å, the ligand is built at an incorrect location.

wrong conformation, and (iii) the intermolecular scale (above 10 Å), where the ligand was built at an incorrect site. Therefore, cases with r.m.s.d. values of equal or less than 1.0 Å were considered to be successful reconstructions. Such a coordinate error should also be (almost) within the radius of convergence of modern refinement programs.

In cases where several identical ligands are present in the PDB entry, the building process (both selection of the density cluster and reconstruction of the ligand) is assessed regardless of which copy of the particular ligand is ‘chosen’ by the software. If, for example, a protein is complexed with three ATP molecules, the reconstruction is declared successful if one of them has been successfully built.

### 3. Results and discussion

#### 3.1. Quality and quantity

The success of any ligand-building procedure strongly depends on the quality of the density map of the ligand looked for. A number of factors affecting this (resolution of the data, ligand size and its ADPs) are discussed in the following sections. Here, we address the point of how good the electron density is overall. Clearly, for a ligand that is hardly bound at all, an attempt to find it would very much seem like Charles Darwin’s blind man looking for a black cat in a dark room when the cat is not actually there. As a measure of the overall quality of the blob of the ligand density, we used a real-space map correlation coefficient (RSMCC) of the density calculated from the deposited reference ligand to the difference electron-density map computed from the X-ray data. The ‘calculated’ density was computed using the coordinates, ADP and atomic types of the reference ligand. An additional artificial exponential term was added to model series terminations. RSMCC was computed on all density grid points that were within 1.9 Å distance of ligand atoms. Fig. 3 displays the distribution of RSMCC for the total test set of single largest and fully occupied ligand-building cases.

The values of RSMCC have a sharp peak at around 0.85–0.95 corresponding to the structures where the ligand is well defined in the density. There is a long tail to the left, some possible reasons for which are elaborated in the subsequent sections. To give a visual impression on how RSMCC relates to the map quality, Fig. 4 presents an example with the PDB entries 1in7 and 1hw8, both refined at a resolution of 2.0 Å and containing a bound adenosinediphosphate molecule. While the density for RSMCC of 93% is very clear, that for RSMCC of 69% lacks experimental support and is difficult to interpret, particularly for the pyrophosphate part of the ligand.

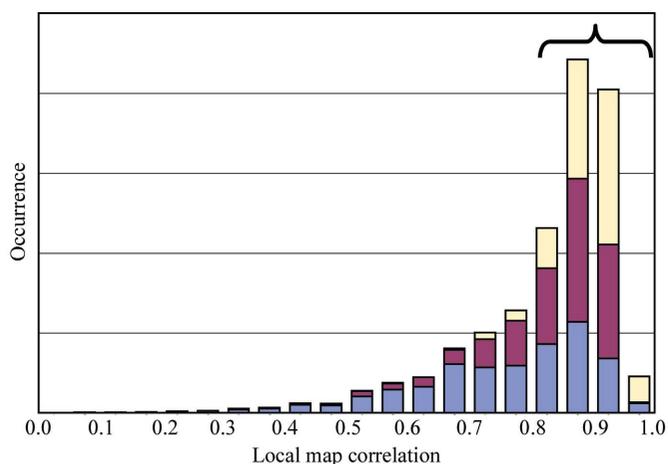
#### 3.2. Overall results

In a small fraction of cases (5–6%) the ligand-building tasks did not proceed to the end, largely owing to the remaining inconsistencies in the PDB file or, sometimes, software crashes. These cases are not taken into account in the subsequent statistical analysis.

An overall assessment of the performance of the procedure along the pipeline of building steps is shown in Fig. 5. Large ligands are built in 27% of cases and small ones in 31% of cases.

For 'good' ligands, however, which we arbitrarily define as those with a size of between ten and 40 non-H atoms and a real-space map correlation coefficient of above 80%, the success rate approaches about half of the cases. The subset of 'good' ligands represents those which are well defined in the density.

We consider separately the performance of each of the two main parts of the ligand-building procedure: the location of



**Figure 3** Distribution of the local map correlation for the test sample of 3369 single largest fully occupied ligands. Those belonging to a 'good' ligand subset are indicated by a curly bracket. The colour coding indicating ligand-building results is the same as in Fig. 5.

the binding site and, given the correct binding site, the construction of the ligand. The overall success rate mentioned above clearly depends on the success of both these two steps.

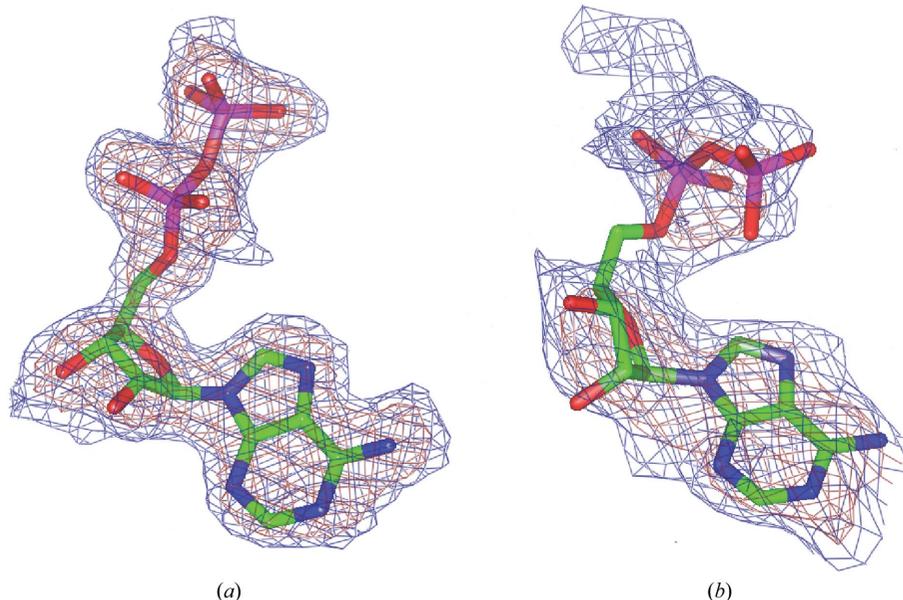
The identification of the ligand-binding site is evaluated on the basis of 3321 entries (*i.e.* there were 48 inconsistencies among the initial 3369 cases). The site has been successfully identified in 64% of the total cases and for 82% for 'good' ligands. This indicates that consideration of the ligand-binding site as a continuous blob of difference density and the implemented method for choosing the appropriate density threshold generally work well. The step of the actual construction of the ligand is based on 2288 cases. The success rate for this step is about 44% overall and is 52% for 'good' ligands. As can be seen from Fig. 3, successful site identification can be expected for ligands with an RSMCC of well above 0.5 and successful ligand construction for ligands with an RSMCC of above 0.7. The sections below present an elaborate discussion on the various factors that affect both steps in the ligand-building procedure in *ARP/wARP*.

### 3.3. Dependence on the ligand size

The location of the binding site and the construction of the ligand molecule are differently affected by the size of the ligand (Fig. 6). Consistently with the intrinsic structure of the algorithm, it is easier to locate the binding sites for larger ligands, while smaller ligands are simpler to build. One reason for this is the fact that the site identification considers all possible clusters in the density by their volume. Conversely, the construction step takes into account all interpretations of a single selected cluster in terms of a ligand molecule.

The success of identifying the location of a ligand does not merely depend on ligand size alone. A major underlying factor is how the correct density cluster competes with other clusters during the process of its selection. These other clusters may originate from other ligands, which (if present) leave their imprint in the electron density, solvent molecules that have been removed from the PDB entry before calculation of the map or noise arising from a possibly incomplete model, errors in phases or observed data, *etc.*

Small ligands, *e.g.* glycerol, sulfate or phosphate, are harder to distinguish from other features in the difference density map and only in about 40% of the cases could their location be correctly determined. However, once located, such small ligands could be very successfully constructed with a high success rate. Medium-sized ligands (20 or more non-H atoms) have larger density clusters that almost always dominate over competing density blobs.



**Figure 4** Examples of density maps with different values of the real-space map correlation coefficient (RSMCC). The ligand adenosinediphosphate is shown as from the PDB entries 1in7 with an RSMCC of 93% (*a*) and 1hw8 with an RSMCC of 69% (*b*). The maps are contoured at  $0.06 \text{ e } \text{Å}^{-3}$  (blue) and  $0.15 \text{ e } \text{Å}^{-3}$  (brown).

This simplifies the task and results in a higher success rate for the location of their site. However, a ligand molecule always has flexible parts (*e.g.* aliphatic chains connecting rigid groups) that may have poorer electron density. The larger the molecule, the more pronounced this effect is, since more such flexible parts may be contained in a ligand. Particularly for large ligands (more than 50 atoms), their location by merely varying the density threshold may become problematic.

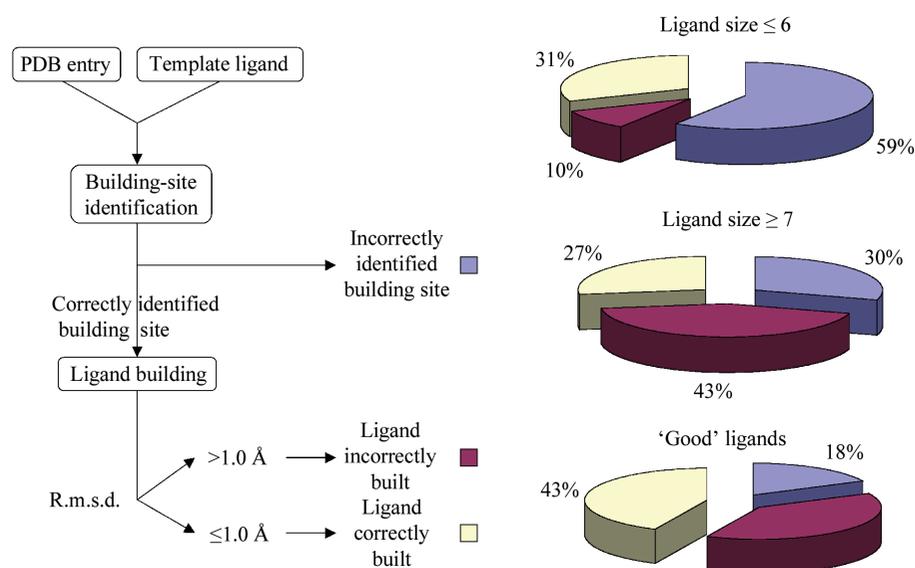
Given the identified site, the success in the construction of the ligand *per se* drops rapidly with increasing ligand size (Fig. 6). The reason lies in the complexity of the label-swapping task when  $N$  atom labels are to be assigned to  $M$  selected cluster grid nodes.  $M$  is on average three times higher than  $N$  and the total number of combinatorial permutations (the upper bound of the number of possible models in the absence of stereochemical information) is thus approximated

by  $(3N)/(2N)!$ . An increasing number of ligand atoms leads to an explosive growth in the number of possibilities for label assignments: while a ten-atom ligand would require checking about  $10^{14}$  assignments in an exhaustive search, a 20-atom ligand would require  $10^{34}$  checks. Limiting the number of possibilities *via* a branch-and-bound approach (Zwart *et al.*, 2004) reduces the computational complexity to about  $3N \times 3^N$ ; the above quoted numbers of possibilities for ten-atom and 20-atom ligands then become  $10^6$  and  $10^{11}$ , respectively. Although this makes the problem computationally tractable, the outcome of the iterative graph search very much depends on the correctness of its early steps. The larger the partial model built at some iteration, the smaller the probability that its possibly incorrect label assignments can be put right by preferring another combinatorial choice. If the ligand to be constructed exceeds a size of 40 atoms, the accumulated misassignments in the construction reduce the success rate considerably.

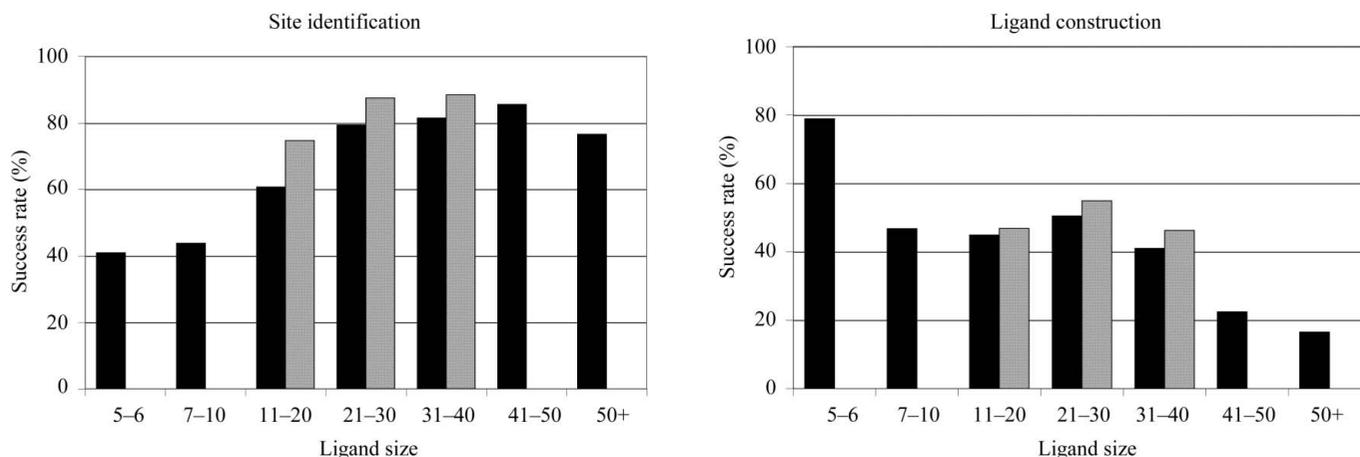
The dependence of ligand building on the ligand size is similar for the total test set and for the subset of 'good' ligands. This seems to reflect the fact that the quality of the density map in the binding site can vary equally well for ligands of all sizes. Both small and large ligands can either be well or poorly bound to the protein, subject to numerous factors of chemical and physical origin.

### 3.4. Dependence on the resolution of the diffraction data

The effect of the resolution of the diffraction data used for the refinement and subsequent construction of the density maps is shown in Fig. 7. The resolution barely influences the success of the identification of the ligand-binding site. This is to be expected, since the volume corresponding to the ligand-



**Figure 5** Flowchart and results of automatic ligand building for 2663 entries with ligands that contain more than six non-H atoms, 706 small ligands with five or six atoms and 1194 'good' ligands. The sites for large ligands are easier to locate, but it is more difficult to construct them.



**Figure 6** Effect of the size of the ligand on the efficiency of the procedure. The total test set is in a darker colour; 'good' ligands are in a lighter colour.

density cluster remains almost unaffected by series-termination effects. However, at very high resolution (1.0 Å and higher) the chances of correct identification of the binding site are lowered. The underlying reason lies in the fact that at such a resolution the atomic features in the density and the fluctuation of the density height along interatomic bonds are very strong. For a value of  $\rho_{\text{thres}}$  higher than about  $1\sigma$  the ligand density is almost always broken into several disconnected small pieces. At a lower threshold the ligand-density cluster is solid but its volume may be of similar size to the volume of density clusters that are located elsewhere and originate from noise, excluded solvent *etc.* This complicates the search of the binding site immensely. In addition, owing to the scarceness of our test set in the resolution range higher than 1.0 Å, these structures have not been taken into account.

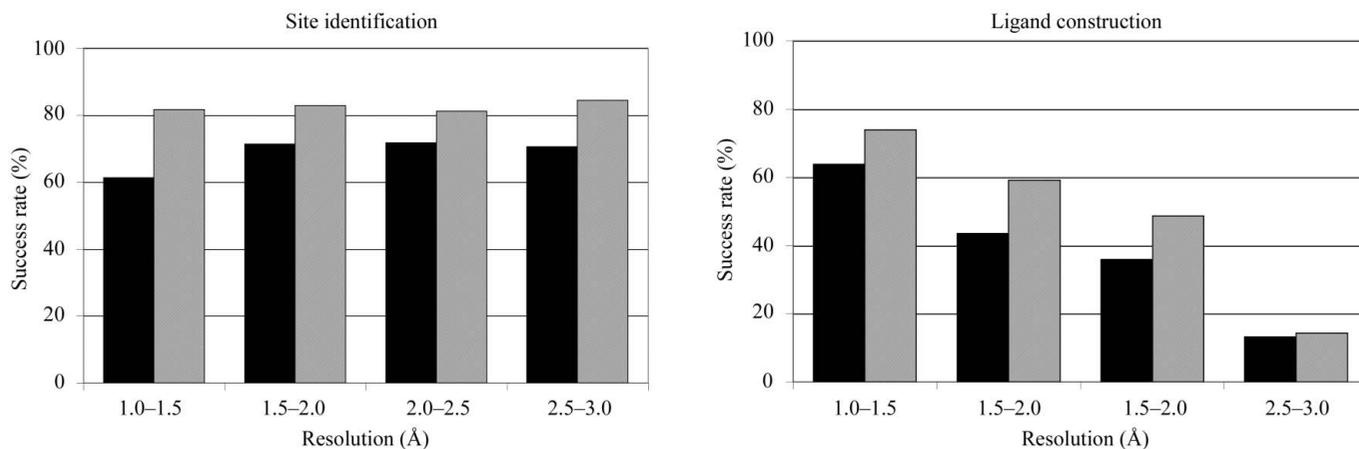
In contrast to site identification, the performance of the ligand construction is affected by the resolution: the lower it is, the lower the success rate (Fig. 7). This could be seen as a direct consequence of the atomicity approach to the description of the electron density which is employed at many steps in *ARP/wARP* (Lamzin & Wilson, 1993). Lower resolution

smoothens the electron density and the discriminative power of the density feature reduces dramatically. Moreover, the shape of the cluster becomes increasingly featureless so that positional ambiguities in the location of sparse grid points cannot be effectively resolved and the stereochemical features used in the ligand-building scoring function become weak. It could thus be concluded that the power of the presented ligand-building procedure does not extend much beyond 2.5 Å.

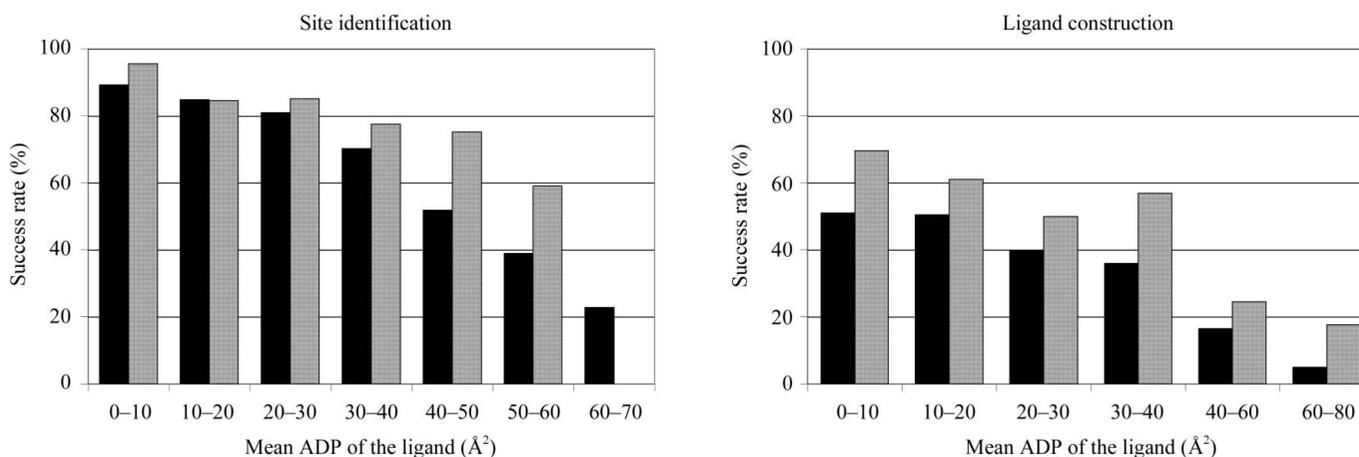
Comparing the dependencies of the total test set and of the subset of ‘good’ ligands does not reveal qualitative differences, as described above in §3.2. A protein–ligand complex may display a wide spectrum of diffraction properties and at any resolution it is equally likely overall to have either well or poorly ordered ligands.

### 3.5. Dependence on the atomic displacement parameters

The disorder of the atoms in the ligand molecule can be estimated by a variety of metrics. We used the average ADP (‘ligand *B* factor’) taken from the deposited structures and



**Figure 7** Effect of resolution of the X-ray data on the efficiency of the procedure. The total test set is in a darker colour; ‘good’ ligands are in a lighter colour.



**Figure 8** Effect of the average ligand atomic displacement parameter on the efficiency of the procedure. The total test set is in a darker colour; ‘good’ ligands are in a lighter colour.

observed that its increase negatively affects both the identification of the ligand site and the ligand construction (Fig. 8). Particularly high ligand  $B$  factors (above  $40 \text{ \AA}^2$ ) bring the performance below the average for the total set. It is understandable that sites with very mobile poorly ordered (and therefore hardly present) ligands are more problematic to identify. In such sites the electron-density cluster is spread much more widely and flatly than for ordered ligands, where it peaks around atomic positions. As a result, given the same volume, the correct cluster becomes indistinguishable from competing density features: other ligands, solvent or noise. Indeed, the cases of 'good' ligands do not show such a sharp falloff with the increase of the ligand  $B$  factor. Concurrently, these are largely the ligands with higher mean  $B$  factor, which have a poorer real-space map correlation coefficient and did not qualify for our subset of 'good' ligands.

The effect of the ADPs on the actual ligand construction (Fig. 8) is similar, but has a slightly different reason: the electron density is smeared over a wider area around the atoms of a ligand, as is the case on lowering the resolution of the data. With a flatter density distribution in a cluster and the reduction of geometric distinctness, during their construction the scores for correct and incorrect partial models lose a good part of their mutual contrast.

### 3.6. Dependence on partial disorder of the ligand

In addition to the mean value of ADPs, we investigated the effect of partial ligand disorder using the ratio of the standard deviation of ADPs within the reference ligand over their mean value as an indicator ( $D$ ),

$$D = \frac{\left[ \sum_{i=1,N} B_i^2 - \frac{1}{N} \left( \sum_{i=1,N} B_i \right)^2 \right]^{1/2}}{\sum_{i=1,N} B_i}$$

The dependence on this is displayed in Fig. 9. The site-identification step is essentially unaffected by partially disordered ligand atoms, except in cases where the value of  $D$  is close to

zero (the first bin in Fig. 9). Although this may look surprising at a first glance, these largely represent cases where the ligand ADPs were set to the same value and possibly not fully refined. This might have been caused by poorly defined electron density, for example. The other reason is related to the way the ADPs are refined. Typically, restraints are used to make the  $B$  factors of neighbouring atoms similar. The degree of similarity is expressed in absolute rather than relative units. Thus, a model with high mean ADP would have a lower value of  $D$  compared with a model with low ADPs. This may lead to a poorer description of the ligand binding than in the PDB and result in lower map correlation coefficients.

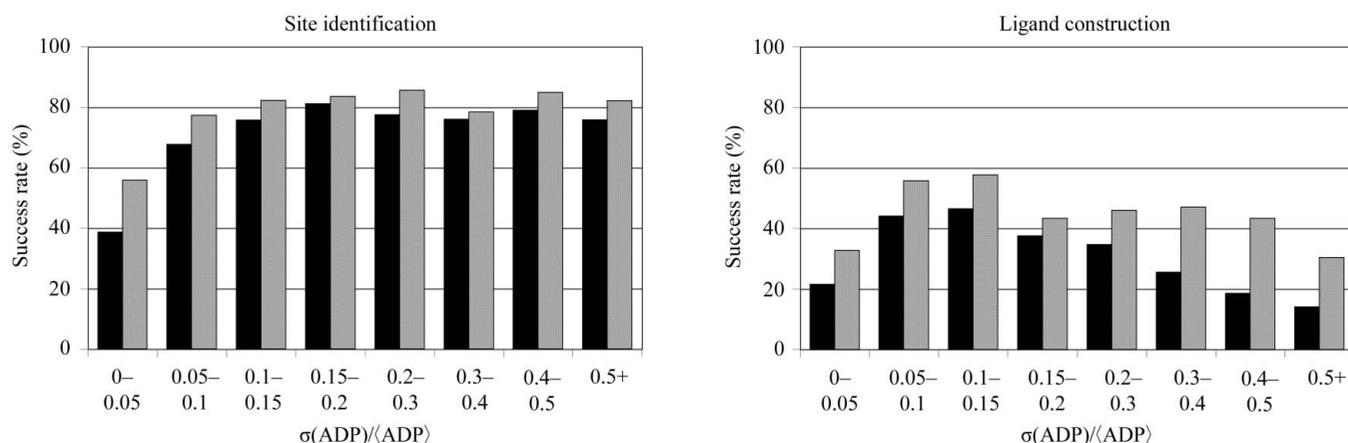
In contrast to the identification of the site, the ligand construction considerably suffers from partial disorder. For both high values of  $D$  (reflecting real partial disorder) and low values (effect of restraints or a lack of refinement), the resolving power of shape features and stereochemical metrics is reduced.

Building partially ordered ligands is a complex scientific task and some means of approaching it are discussed in §4.

### 3.7. Building a ligand that is not the largest

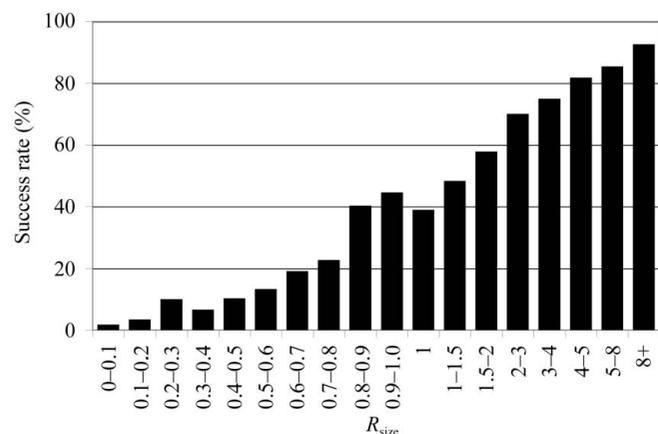
As described above, the building procedure in *ARP/wARP* version 6.1 has been designed to build one (the largest) ligand at a time. Nevertheless, here we briefly overview the results of a suboptimal use of the software in which the reconstruction has been attempted on any ligand, regardless of whether a larger ligand might have been present. The analysis is limited to the site-identification stage.

Based on the design of the procedure, when a mixture of ligand types is present, ligand-binding sites are most successfully identified for the largest ligand type. This behaviour is captured in Fig. 10. The success of identifying the location of a ligand is strongly correlated with the ratio  $R_{\text{size}}$ , which we define as the ratio of the size of the ligand to be built to the size of the largest of the remaining ligands bound to the protein. For values of  $R_{\text{size}}$  that are lower than one, *i.e.* an attempt to build a ligand that is not the largest in the structure,



**Figure 9**

The dependence on the partial disorder as judged by the ratio of the standard deviation of ligand ADPs to their mean value. The total test set is in a darker colour; 'good' ligands are in a lighter colour.



**Figure 10**  
Efficiency of the identification of the binding site in the presence of other smaller ( $R_{size} > 1$ ) or larger ( $R_{size} < 1$ ) ligands.

the likelihood of correct site identification rapidly approaches zero. At the same time, in a considerable number of cases the location is still correctly identified, particularly when  $R_{size}$  is within the range 0.8–1.0. The reason for this probably lies in the search of the optimal density threshold  $\rho_{thres}$ , where the cluster for a smaller ligand may survive an increase of the contour level at nearly the same volume, whereas the density for a competing larger entity may break into unconnected pieces. Fig. 10 also shows that even for the search of the largest ligands, the success rate for the correct identification of the binding site is still dependent on the size contrast between competing ligands. Given a noisy electron-density map, the clusters corresponding to similarly sized but different ligands may become indistinguishable just by their volume.

#### 4. Concluding remarks

An advantage of the ligand-building module in *ARP/wARP* is its ease of use *via* a *CCP4i*-type graphical user interface (Potterton *et al.*, 2003) or *via* the command line where a user can setup his/her own pipeline for repeated runs or fast execution of building tasks. In this way, it lends itself to drug-development and medium-throughput efforts when a list of compounds are to be screened against the same protein.

The casting of ligand construction into the framework of an optimization problem and the quasi-exhaustive graph-search algorithm allow local modelling decisions to be dealt with in a very efficient and yet general manner while consulting the result at a more global level, keeping various models in memory for comparison. The presented approach is prototypical in nature owing to its clear distribution of subtasks, namely site identification and ligand construction. The method resembles what a human would do: identify the site where to construct the molecule considering all plausible density blobs and, having done that, construct the model of the ligand judging from the shape of the density at the chosen location.

In the current implementation, only the volume of the electron-density clusters is used for selection of the optimal density threshold and the identification of ligand sites. This leads to problems for multi-ligand cases, very high resolution

data, partial disorder or a high level of noise in the map. Learning from the obtained results, we envision various ways in which the ligand-site identification may be improved. One is the use of the ligand's density-shape information, *i.e.* how well the experimental density matches stereochemical expectations. Knowledge of the host protein can be exploited to discriminate binding scenarios at several plausible binding sites taking the chemical (hydrogen-bonding) environment into account. Possibly, a number of top-ranked cluster candidates could be tried for a fast ligand construction and the final choice made upon the values of the obtained RSMCC. High resolution is ideal for ligand construction since the density height becomes the dominant feature in the scoring. However, performance currently breaks down with data to a resolution lower than about 2.5 Å owing to the lack of atomicity. The use of a sparse cluster representation simplifies the general problem of the combinatorial tractability of the graph-search technique. However, the sparsing should still provide a sufficiently good parameterization of the density and accurate enough seed points for ligand construction; these also depend on the atomic features of the map. The basic idea of label swapping may be extended to incorporate all map grid points within a selected density cluster as well as other parameters of importance, including the presence of rigid groups and preferred conformations. Thus, a more sophisticated use of prior information may prove helpful, especially in providing better likelihood targets for the construction step.

An issue which is not at all taken into account by the ligand-building module in *ARP/wARP* version 6.1 is partial occupancy and particularly high local disorder of the ligand to be built. In the analysed PDB set we skipped almost 10% of ligand-building tasks in which the ligand was explicitly modelled with partial occupancy. This was also the reason to separate, on the basis of the values of the real-space correlation coefficient, and consider separately 70% of the tasks as 'good' ligand cases. Partial occupancy or a complete absence of a part of the ligand directly affects the shape of the density cluster. If, despite this difficulty, a density cluster is somehow selected correctly (or already known), it may still not have the correct shape.

The ligand-building procedure evaluated in this paper is a sequence of successive steps and is therefore reminiscent of a software pipeline. Its efficiency is thus mostly affected by the weakest element in the chain. For example, while larger ligands are easier to locate but harder to build, the highest success rate can be observed for ligands within the range 20–30 atoms. This factor of the chain efficiency reappears in the multi-ligand case when the largest ligand should be built first in order to maximize the overall success rate.

Finally, we note that the aim of this study was to help in making the ligand-building software more advanced and to assess its performance on the example of *ARP/wARP* version 6.1 subject to various working conditions close to the standard use of the software. Ongoing research will address a number of shortcomings that were pointed out by the study detailed in this paper. The main benefit of the presented large-scale test of the software is to give representative results, where repro-

ducibility in a single instance is provided as a probability estimated on the safe grounds of a large number of test cases. Terwilliger *et al.* (2006) subject the evaluation of their method to the same paradigm. Valid conclusions as to the efficiency of today's complex model-building software can only be drawn if more cases shed further light on every instance of a particular method.

The authors would like to thank Gerard Kleywegt for his contribution and help with the EDS and HIC-Up databases, Diederick de Vries for help with the SQL database and Garib Murshudov for fruitful discussions regarding the use of *REFMAC5*. The work was supported by the NIH R01 GM62612-01 grant in a form of a fellowship to GGL.

## References

- Aishima, J., Russel, D. S., Guibas, L. J., Adams, P. D. & Brunger, A. T. (2005). *Acta Cryst.* **D61**, 1354–1363.
- Arfken, G. (1985). *Mathematical Methods for Physicists*, 3rd ed, pp. 964–965. Orlando, FL, USA: Academic Press.
- Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.
- Diller, D. J., Pohl, E., Redinbo, R. R., Hovey, B. T. & Hol, W. G. J. (1999). *Proteins*, **36**, 512–525.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126–2132.
- Gasteiger, J., Rudolph, C. & Sadowski, J. (1990). *Tetrahedron Comput. Methods*, **3**, 537–547.
- Golovin, A., Dimitropoulos, D., Oldfield, T. J., Rachedi, A. & Henrick, K. (2005). *Proteins*, **58**, 190–199.
- Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* **D60**, 2240–2249.
- Kleywegt, G. J. & Jones, T. A. (1998). *Acta Cryst.* **D54**, 1119–1131.
- Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* **D49**, 129–147.
- Morris, R. J., Najmanovich, R. J., Kahraman, A. & Thornton, J. M. (2005). *Bioinformatics*, **21**, 2347–2355.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.
- Oldfield, T. J. (2001). *Acta Cryst.* **D57**, 696–705.
- Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* **D59**, 1131–1137.
- Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B. & Kuriyan, J. (2000). *Science*, **289**, 1938–1942.
- Terwilliger, T. C., Klei, H., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2006). *Acta Cryst.* **D62**, 915–922.
- Zwart, P. H., Langer, G. G. & Lamzin, V. S. (2004). *Acta Cryst.* **D60**, 2230–2239.